OPEN ACCESS Freely available online

PLOS ONE

100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox

Francisco J. Valverde-Albacete[1]*, **Carmen Peláez-Moreno**[2]

1 Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, Madrid, Spain, **2** Signal Theory and Communications Department, University Carlos III Madrid, Madrid, Spain

Abstract

The most widely spread measure of performance, accuracy, suffers from a paradox: predictive models with a given level of accuracy may have greater predictive power than models with higher accuracy. Despite optimizing classification error rate, high accuracy models may fail to capture crucial information transfer in the classification task. We present evidence of this behavior by means of a combinatorial analysis where every possible contingency matrix of 2, 3 and 4 classes classifiers are depicted on the entropy triangle, a more reliable information-theoretic tool for classification assessment. Motivated by this, we develop from first principles a measure of classification performance that takes into consideration the information learned by classifiers. We are then able to obtain the entropy-modulated accuracy (EMA), a pessimistic estimate of the expected accuracy with the influence of the input distribution factored out, and the normalized information transfer factor (NIT), a measure of how efficient is the transmission of information from the input to the output set of classes. The EMA is a more natural measure of classification performance than accuracy when the heuristic to maximize is the transfer of information through the classifier instead of classification error count. The NIT factor measures the effectiveness of the learning process in classifiers and also makes it harder for them to "cheat" using techniques like specialization, while also promoting the interpretability of results. Their use is demonstrated in a mind reading task competition that aims at decoding the identity of a video stimulus based on magnetoencephalography recordings. We show how the EMA and the NIT factor reject rankings based in accuracy, choosing more meaningful and interpretable classifiers.

Citation: Valverde-Albacete FJ, Peláez-Moreno C (2014) 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox. PLoS ONE 9(1): e84217. doi:10.1371/journal.pone.0084217

Editor: Matteo G. A. Paris, Università degli Studi di Milano (University of Milan), Italy

Received July 22, 2013; **Accepted** November 13, 2013; **Published** January 10, 2014

Funding: Francisco José Valverde-Albacete has been partially supported by EU FP7 project LiMoSINe (contract 288024): www.limosine-project.eu Carmen Peláez Moreno has been partially supported by the Spanish Government-Comisión Interministerial de Ciencia y Tecnología project TEC2011–26807. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: fva@lsi.uned.es

Introduction

Classification is an ubiquitous task in Science, Technology and the Humanities [1]. Usage ranges from diagnosing diseases [2] or the status of tumors using gene expression data [3] to the actual classification of tumor classes [4]; from analyzing human performance in perceptual tasks [5] to analyzing that of automated remote sensors [6] or automatic speech recognition machines [7]. If follows that the assessment of the performance of classification processes is of paramount importance for Scientific, Technological and Societal reasons [1,8–10].

To set the theoretical backdrop for our discussion, consider a set of k *prior, instance or true classes* $\{x_1, \ldots x_k\}$ and a discrete random variable X distributed according to a *prior class distribution* P_X. Consider also a set of N *instances or patterns*, each belonging to only one of those classes, but we do not know precisely which. A *classification* is a process whereby each of those instances is assigned to one among a set of m *decision or predicted classes* $\{y_1 \ldots y_k\}$ generating a discrete random variable Y distributed according to a *posterior class distribution*, P_Y, so that the joint events of this classification process consist of "presenting one instance of an input class $X = x_i$ for classification and deciding the output class to be $Y = y_j$".

To measure the performance of the classification process we use its *confusion matrix*, a special contingency table C_{XY} counting the occurrences of the joint events. Usually, the maximum likelihood estimate of the joint probability $P_{XY} \approx C_{XY}/N$ is used as summary data. Figure 1 represents two such contingency matrices for a *brain decoding* or *mind reading* task consisting in automatically identifying the class of video stimulus shown to the subjects based on magnetoencephalography (MEG) data. Five different types of stimuli were presented: the first three ones (x_1, x_2 and x_3) belonging to the category of *short* clips (6–26 s. long) and the last two (x_4 and x_5) to the category of *long* clips (approximately 10 min. long).

Performance evaluation takes the form of the exploratory analysis of this confusion matrix or joint distribution. For instance, the *de facto* standard for performance *visualization* for binary—that is, two-class—classification is the Receiver-Operating-Characteristic (ROC) [11], but its generalization to higher class numbers is not as effective. We have argued elsewhere that the De Finetti entropy triangle (ET) [12] is a better tool to analyze classifier performance, with a solid information-theoretical basis, and not

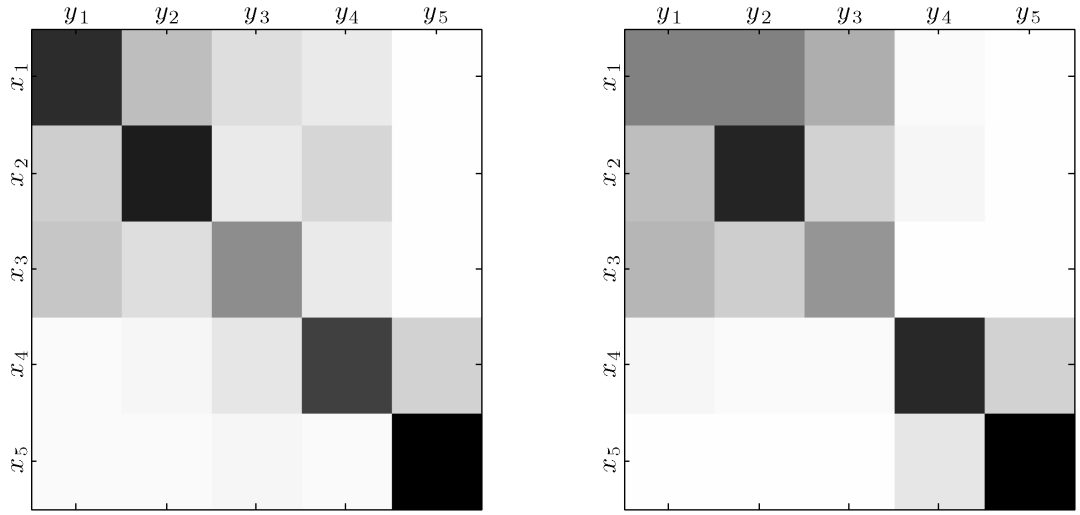

Figure 1. Heatmap of the best classifiers of the MEG mind reading competition [23] according to accuracy (left) and the EMA and the NIT factor (right) criteria. Rows correspond to stimulus $X = x_i$ and columns to the decision $Y = y_j$ or response. Darker hues correlate with higher joint probability P_{XY}. The heat map on the left reveals that the best classifier according to accuracy does not capture the fact that stimuli x_1, x_2 and x_3 belong to a particular category whilst x_4 and x_5 belong to another. A$k = 2, a(P_{XY}) \in \{0.50, 0.75, 0.88, 0.94, 0.97, 0.98, 1.0\}$ B$k = 3, a(P_{XY}) \in \{0.33, 0.67, 0.83, 0.94, 1.0\}$ C$k = 4, a(P_{XY}) \in \{0.25, 0.63, 0.81, 0.94, 1.0\}$
doi:10.1371/journal.pone.0084217.g001

plagued with the problems of the ROC—see *sec:mms: sec:entropy-triangle*. In any case, neither device provides a *single* figure-of-merit or performance measure to compare systems, a practice cherished by researchers.

As a single figure-of-merit, by far the most widespread performance criterion used is *accuracy*, defined as the fraction of correctly classified instances, $a_{P_{XY}} = trace(P_{XY})$. This is probably due to its easy and intuitive nature, despite many reasons *not* to do so [13]. In [14], this and many other performance measures were examined in the context of several machine learning tasks, but inconclusive results as to their fitness of purpose were reached. However, the comparison made evident that accuracy was one of the measures that possessed the least number of invariants with respect to changes in confusion matrix entries, a detrimental quality. An earlier paper [15] had already argued for the factoring *out* of the influence of prior class distributions on similar measures.

It is now acknowledged that *high accuracy is not necessarily an indicator of high classifier performance* and therein lies the *accuracy paradox* [16–18]. For instance, in a predictive classification setting, predictive models with a given (lower) level of accuracy may have greater predictive power than models with higher accuracy. This deleterious feature is explained in-depth in Section *sec:crit-accur-using*. In particular, if a single class contains most of the data, a *majority classifier* that assigns all input cases to this majority class (the one concentrating the probability mass of P_X) would produce an accurate result. Highly *imbalanced* or *skewed* training data is very commonly encountered in samples taken from natural phenomena. Moreover, the classes' distributions of the samples do not necessarily reflect the distributions in the whole population since most of the times the samples are gathered in very controlled conditions. This skewness in the data hinders the capability of statistical models to predict the behavior of the phenomena being modeled and data balancing strategies are then advisable [19].

In this paper, we claim that performance measures based in the statistical information transfer from X to Y may be better measures for classification if *predictive classification error is not the paramount performance criterion*. This is the case, for example, of classifiers not used to make final decisions but, instead designed to be components of more complex diagnostic systems (as in [19]) or when the conditions in the experimentation stage during which the data is collected do not hold in the deployment stage, as mentioned before. For this purpose, in Section *sec:perpl-its-prop* we establish the basis of our analysis in the propagation of *perplexity*—the effective number of classes a classifier sees—a concept that is directly related to accuracy.

In Section *sec:perf-meas-based* we use the *remaining input perplexity* $k_{X|Y}$ to claim that the *entropy-modulated accuracy (EMA)*, defined in (3), is a better measure of classifier performance than accuracy for several reasons: it is well-grounded in information-theoretical terms, it provides an intuitive interpretation of the statistical learning process as the transfer of the information from the phenomena that are being modeled over a virtual channel, it factors out the influence of the input and output class distributions, it is invariant to permutations in the columns of the confusion matrix enabling the identification of cross-labeling errors common in unsupervised learning methods, and it is a pessimistic estimate of accuracy. For the same reasons, the *normalized information transfer factor (NIT factor)*, defined as in (5), adds to some of the previous advantages the fact that it is capable of assessing the effectiveness of the learning process in the classifier, it is co-variant with expected mutual information (MI) [20], and contra-variant with the variation of information [21].

In *sec:example-use*, we suggest how to apply these metrics to a classification task, instantiating the process for a mind-reading challenge using multi-classification on magnetoencephalography signals, that shows one clear instance where ranking by EMA and NIT factor provides a more interpretable classifier than accuracy-based ranking. We provide further evidence, examples and a comparison with other metrics in *File S1*. The paper is closed with a *sec:discussion* where we also compare EMA and the NIT factor with two previously proposed measures for classification assessment and show the superiority of our proposal.

Results

A critique of accuracy using information-theoretic principles

To assess the theoretical adequacy of accuracy, we generated some samples of the space of joint count distributions for k input and output classes and N instances of classification with a prescribed accuracy (see Section *sec:datasets* for the details). Then, their entropy decomposition was calculated and plotted in the ET (see Section *sec:mms*, *sec:entropy-triangle*). Figure 2 presents the cases $k=2$ with $N=100$, $k=3$ with $N=18$ and $k=4$ with $N=16$.

A number of observations can be gleaned from this figure:

- *Matrices of a particular accuracy level are interspersed with those of many other accuracy levels*. This phenomenon is the more prevalent the lower the accuracy level, although the behavior differs for different k. For $k=2$ interspersing ends for accuracies over 0.75 while for $k\geq 3$ it spreads to the whole range $[1/k,1.0]$.
- *For every prescribed accuracy level, the normalized mutual information ranges in* $[0,1]$, that is, there are matrices with accuracy over $1/k$ transmitting little or no information. This is the case even for high-accuracy matrices, including those with accuracy 1.0.
- Conversely, matrices with different accuracy may exhibit the same normalized mutual information, for instance, check at $2MI'_{P_{XY}}=0.6$.

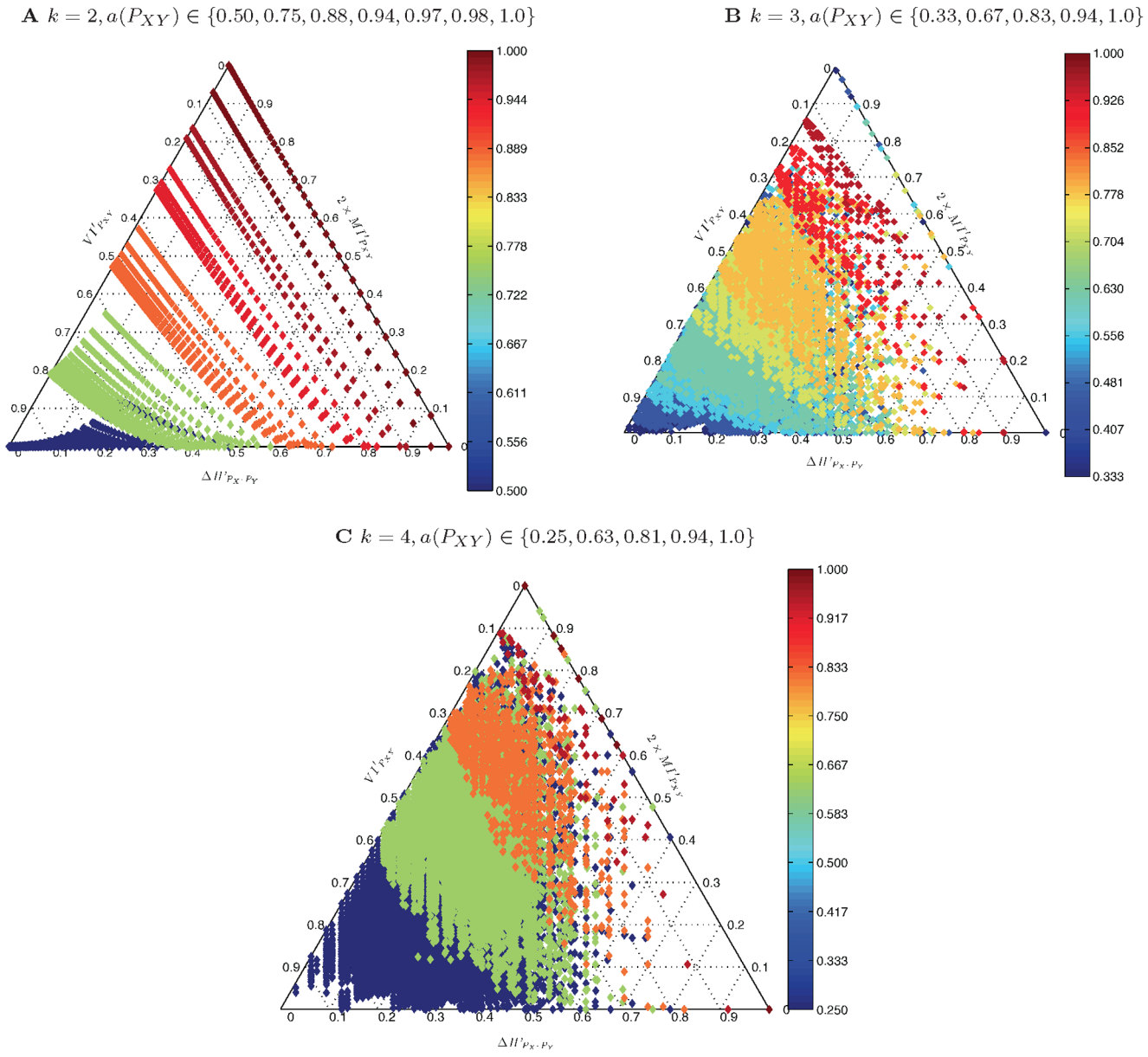

Figure 2. (Color online) Entropy decomposition for square matrices of (A) $k=2$, (B) $k=3$, and (C) $k=4$ (decimated), representing confusion matrices for a classification task at different accuracy levels as described by the right color bar. The interspersing of the plots representing matrices with different accuracies but similar entropies is evident at all levels for $k=3$ and $k=4$ but only for lower levels of accuracy for $k=2$. This entails that accuracy is not a good criterion to judge the flow of information from the input labels to the output labels of a classifier (see text).

doi:10.1371/journal.pone.0084217.g002

- There is an accumulation of distributions with high entropy (low $\Delta H_{P_{XY}}$ values, left side of ET), as predicted by theory [22].

We are driven to conclude that accuracy is not a trustworthy criterion to judge the degree to which a particular classification process transfers information from the input class distribution to the output decision class distribution.

Perplexity and its propagation in multiclass classifiers

The question poses itself whether it is possible to conjoin accuracy and mutual information transfer in a single measure. To provide an affirmative answer to this we first state the hypothesis:

Hypothesis 1. In the absence of information about the items distributed according to a uniform prior class distribution, a classifier is expected to guess correctly $1/k$ of the times.

We will show that the EMA amounts to a 'pessimistic' accuracy estimate according to this hypothesis. For the sake of generality, suppose that the cardinality of the set of atomic events of P_X is k and that of P_Y is m. Classification tasks with uniform input class distributions are often called *balanced* or *unskewed*. Let us denote this uniform input distribution as U_X and accordingly, U_Y will represent a uniform distribution of the outputs. Now H_{U_X} and H_{U_Y} represent the entropies of U_X and U_Y respectively. Then $k=2^{H_{U_X}}$ and $m=2^{H_{U_Y}}$, so k is a measure of the *theoretical perplexity* of a classifier in a balanced task, that is, the number of *possible* events.

By analogy, call $k_X=2^{H_{P_X}}$ and $m_Y=2^{H_{P_Y}}$ the *perplexities* of variables X and Y respectively. They are in fact an estimation of the *effective*—as opposed to the *possible*—number of atomic events behind P_X and P_Y. Note that $1\leq k_X\leq k$ and $1\leq m_Y\leq m$ and that $k_X=k$ ($m_Y=m$) precisely when $P_X=U_X$ ($P_Y=U_Y$). Similarly, $k_X=1$ ($m_Y=1$) when P_X (resp. P_Y) resembles a Kronecker delta function—that is, the input (and output) distribution is utterly skewed towards one class.

If we now define the quotient $\delta_X=\frac{k}{k_x}$ (respectively, $\delta_Y=\frac{m}{m_Y}$) we can see that

$$\delta_X=\frac{k}{k_x}=2^{H_{U_X}-H_{P_X}}=2^{\Delta H_{P_X}}$$

$$(\delta_Y=\frac{m}{m_Y}=2^{H_{U_Y}-H_{P_Y}}=2^{\Delta H_{P_Y}}),$$

where $\Delta H_{P_X}=H_{U_X}-H_{P_X}$ ($\Delta H_{P_Y}=H_{U_Y}-H_{P_Y}$.) We interpret this quantity as the decrement (increment) in perplexity due to the choice of input (output) marginals of P_{XY}.

The most important concept in our discussion is the *information transfer factor* $\mu_{XY}=2^{MI_{P_{XY}}}$: if we introduce two new *remaining perplexities*, $k_{X|Y}=2^{H_{P_{X|Y}}}$ and $m_{Y|X}=2^{H_{P_{Y|X}}}$, from the well-known formulae $MI_{P_{XY}}=H_{P_X}-H_{P_{X|Y}}=H_{P_Y}-H_{P_{Y|X}}$ this crucial quantity can be understood as the perplexity variation of X and Y produced by the subtraction/addition of their mutual information,

$$\mu_{XY}=2^{H_{P_X}-H_{P_{Y|X}}}=\frac{k_X}{k_{X|Y}}$$

$$=2^{H_{P_Y}-H_{P_{Y|X}}}=\frac{m_Y}{m_{Y|X}},$$

hence the name.

It is easy to see that we have completed two different, sequentially related, decompositions of the perplexity of the variables,

$$k=\delta_X\cdot k_X=\delta_X\cdot\mu_{XY}\cdot k_{X|Y} \quad (1)$$

$$m=\delta_Y\cdot m_Y=\delta_Y\cdot\mu_{XY}\cdot m_{Y|X}.$$

This proves that an alternative way of conceptualizing the flow of information from one variable to the other is in terms of increments or decrements of their perplexity instead of the flows of entropies, as depicted in Fig. 3. In fact, the following inequalities can easily be checked,

$$k\geq k_X\geq k_{X|Y}\geq 1 \qquad 1\leq m_{Y|X}\leq m_Y\leq m. \quad (2)$$

Note that analogue decompositions for marginal entropies were introduced in [12], and are here collected as *sec:mms: sec:split-entr-triangle*. We will see next how this conceptualization allows us to devise an alternative to accuracy where the decomposition of equation (1) underlines the preeminence of P_X for assessing performance.

Two performance measures based on perplexity

Consider a confusion matrix for a classifier obtained from N instances of classification pairs. The lowest accuracy is that of a classifier returning a uniform count matrix: the most balanced testing dataset will distribute N/k to each class and a clueless classifier will further redistribute these uniformly to each output class as $N/(km)$ instances. Since the diagonal has $\min(k,m)$ cells, the diagonal sum is

$$\text{trace}(C_{XY})=\sum_{i=1}^{min(k,m)}\frac{N}{km}=\frac{N}{\max(k,m)},$$

whence the accuracy is

$$a(P_{XY})=\frac{\text{trace}(C_{XY})}{N}=\frac{1}{\max(k,m)}=\min(\frac{1}{k},\frac{1}{m}).$$

It is bounded by $\min(\frac{1}{k},\frac{1}{m})\leq a(P_{XY})\leq 1$ and any value smaller than the lower bound is an sure indication that a permutation of the output tags will ensure higher classification accuracy, that is, a better mapping of input to output *class names*.

Consider the perplexity reduction chain of Fig. 3. To the extent that the number of input classes and their distribution is a given—whereas P_Y is a construct of the classifier—we want to concentrate on measuring how well the input class distribution was learned by the training process, that is, in the prior class distribution perplexity reduction of equation (1). Regarding the classifier training algorithm, ΔH_{P_X} is a given and cannot be modified, whereas $MI_{P_{XY}}$ quantifies the amount of *successfully* learned information. More importantly for our purposes, $H_{P_{X|Y}}$ is the amount of information the classifier *failed* to learn. Therefore the EMA appears naturally as a quality measure based in the remaining perplexity of the X variable

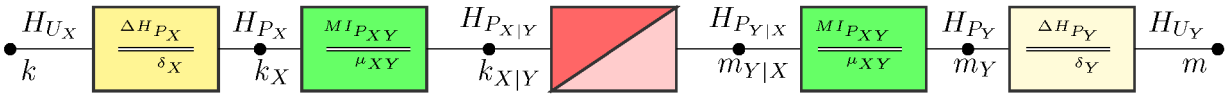

Figure 3. (Color online) Entropy (above) and perplexity (below) decomposition chains for a joint distribution. Left, perplexity reduction in the input (learning) chain; right, perplexity increase in the output chain, related to classifier specialization. The colors refer to those of Fig. 5.(B). The ordering of the boxes is a convention to reveal the prior and posterior natures of the perplexities of class distributions.
doi:10.1371/journal.pone.0084217.g003

$$a'(P_{XY}) = \frac{1}{k_{X|Y}} = \frac{1}{2^{H_{P_{X|Y}}}}, \quad \frac{1}{k} \le a'(P_{XY}) \le 1. \tag{3}$$

Since $H_{P_{X|Y}}$ is the entropy of P_X ignored by the classifier, as per our hypothesis and in the absence of any other source of information *this is the expected performance of the classifier with equivalent (possibly fractional), equally likely $k_{X|Y}$ classes*: the higher this number, the worse the classifier will be.

To illustrate this, notice that when the training process of the classifier has been able to capitalize on all mutual information to leave no remaining perplexity, $k_{X|Y} = 1$, whence $a'(P_{XY}) = 1$. Similarly, $a'(P_{XY}) = \frac{1}{k}$, either because the classifier has utterly failed to capture any information between X and Y, $\mu_{XY} = 1$, or because the entropy of the data was minimal, $\delta_X = k$.

Notice that when the entropy of P_X is not maximal $H_{P_X} \neq H_{U_X} = \log(k)$ then $\delta_X > 1$ whence $k > k_X \geq k_{X|Y}$ and the EMA detects an *artificial* lower bound for (2), $a'(P_{XY}) = 1/k_X$. The artifice here is that this increase does not depend on the training of the classifier but on the prior class distribution. This suggests including a correction into equation (3) to account for the deviation from uniformity in the prior class distribution $\Delta H_{P_X} = H_{U_X} - H_{P_X} \neq 0$ so that

$$q(P_{XY}) = \frac{\delta_X}{k_{X|Y}}, \quad \frac{1}{k} \le q(P_{XY}) \le a'(P_{XY}) \le 1, \tag{4}$$

with $q(P_{XY}) = 1$ when both $\mu_{X|Y} = k$ and $k_X = k$, implying that $\delta_X = 1$ and $k_{X|Y} = 1$. Note that $q(P_{XY}) = a'(P_{XY})$ if and only if $P_X = U_X$. Unlike the case of $a'(P_{XY})$, the eventuality that the data are not uniformly distributed is corrected on $q(P_{XY})$, as $\delta_X \neq 1$ entails $k_X \neq k$. Moreover, the further away from a uniform prior class distribution to the classifier, the worse its upper range bound will be. Eventually, for $k_X = 1$—which implies $k_{X|Y} = 1$ by equation (2) whence $\mu_{XY} = 1$—we have, again, the worst possible value of the measure, $q(P_{XY}) = 1/k$. Notice that in this accuracy-optimal case $a'(P_{XY}) = a(P_{XY}) = 1$, but in an unhelpful way. Essentially, making the input data less random impacts the ability of the classifier to capitalize in mutual information to bind together input and output, and this is registered by the measure. The *normalized information transfer factor* can be rewritten as,

$$q(P_{XY}) = \frac{\mu_{XY}}{k_X} \cdot \frac{k_X}{k} = \frac{\mu_{XY}}{k} = \frac{2^{MI_{P_{XY}}}}{k} \tag{5}$$

Note also that NIT factor does *not* depend directly on the input or output distribution. Conveniently, since the normalized information transfer factor is a monotonic function of normalized mutual information the relative height in the ET offers a visual tool to quickly inspect such effectiveness. Finally, when evaluating a set of systems in the same task, δ_X is constant throughout the evaluation, so $a'(C_i) \propto q(C_i)$, and they offer the same ranking results, easily visualized in the ET.

For the reasons above, we posit the EMA in (3) to measure the performance of classification tasks, and the NIT factor in (4) or (5) to measure the effectiveness of the classifier learning process.

Assessing classifiers with EMA and the NIT factor

In this Section we present an example of how to use the EMA and the NIT factor in automatic classifier evaluation tasks. We consider the case of the MEG mind reading challenge organized by the PASCAL (Pattern Analysis, Statistical modeling and ComputAtional Learning) network [23]. Since accuracy was the "official" evaluation criterion, for comparison purposes Fig. 4.fig: (A) presents the results in the entropy triangle ordered by accuracy as reflected in the coloring of the points. System C_1 at $a(C_1) = 0.680$ was deemed the winner with C_2 close behind at $a(C_2) = 0.632$. In a detail of the dense region of harder competition in Fig. 4.(B) clusters $\{C_4, C_2\}$, $\{C_1, C_3\}$ and $\{C_6, C_5, C_7\}$ are evident. We next suggest a procedure to analyze the classification performance of a population of classifiers:

1. **Use k_X to assess the effective number of classes of the data.** At $k_X = 4.950$ down from $k = 5$, the task is quite balanced, guaranteeing that systems will find it harder to specialize as majority classifiers.
2. **Use EMA to rank classifiers.** Table 1 presents the perplexities, accuracies, the EMA and the NIT factor for the confusion matrices of the classifiers that took part in the task. Ranking $(C_4, C_2, C_1, C_3, C_6, C_5, C_7, C_9, C_8, C_{10})$ suggests itself, aligned with increasing mutual information (right axis). Indeed, after EMA, C_4 should have been the winner of the competition, followed closely by C_2.
3. **Use the ET to individually assess each classifier.** From the ET diagram it is evident that those classifiers with highest mutual information and accuracy—the first seven classifiers—are not specialized while classifier 10, and, to a lesser extent, 8 and 9 are. The worst classifier is barely above random at $q(C_{10}) = 0.206$.
4. **Use the NIT factor to assess whether the population of classifiers has solved the task.** Overall, for the top ranked classifier we have $q(C_4) = 0.407$, showing that the task has indeed not been effectively solved by the participants, either individually or collectively.

The result of this process is an assessment of a (population of) classifiers, whereby one may discuss the advantages of EMA and NIT factor vis–vis other performance measures, for instance, accuracy. Further examples of using this procedure to evaluate classification tasks can be found in the *File S1*.

EMA and NIT factor vs. Accuracy. The authors of the report on the MEG Mind Reading challenge attempted an analysis of the ranking results and specifically compare classifier C_1 to C_4 since the heat map of the latter seems to be "cleaner"

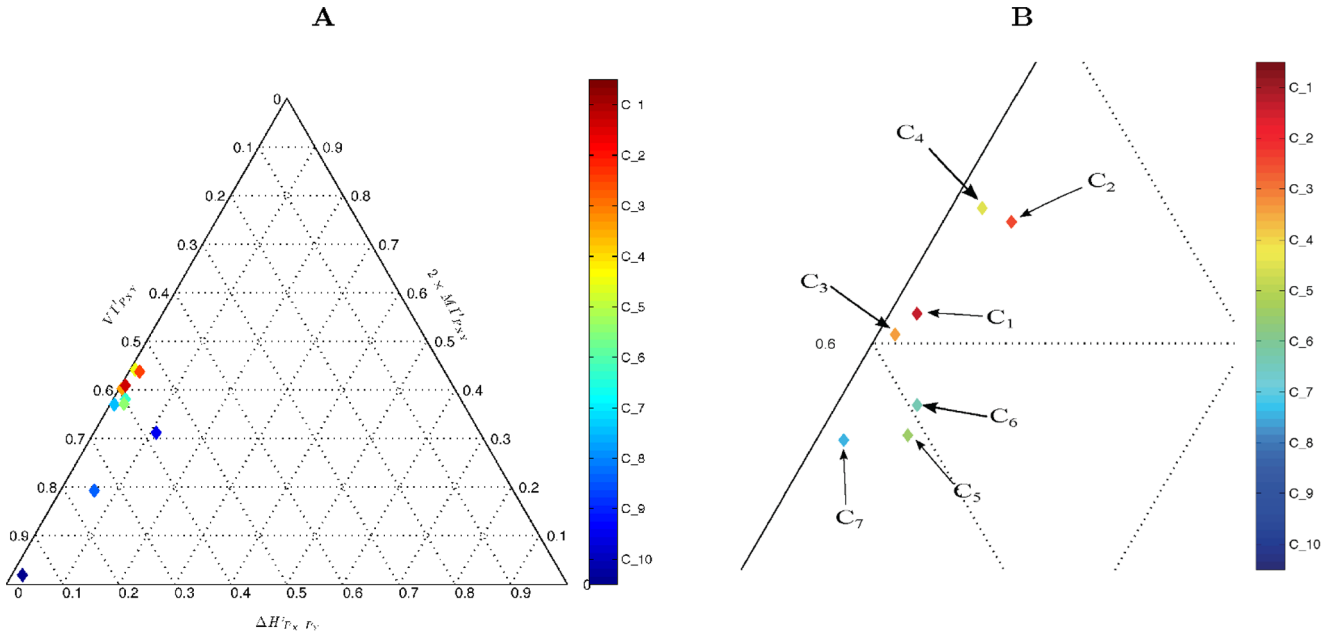

Figure 4. (Color online) Entropy triangle for the MEG mind Reading data ordered after accuracy (A) and a detail of the participants of higher accuracy (B). The ranking following accuracy is at odds with the EMA and the NIT factor ranking based in mutual information (height, right scale of triangle). The detail in (B) shows that participant C_4, closely followed by C_2 should have been ranked first after this criterion.
doi:10.1371/journal.pone.0084217.g004

[23] (see Fig. 4 with the heat map of C_1 (left) to C_4 (right)). For them, classifier C_1 essentially came out first because it used the "learning capacity" of its technique to improve classification error while C_4 used the capacity to better distinguish the two categories of classes present in the task (with stimuli x_1 to x_3 belonging to a first category whilst x_4 and x_5, to another) but was worse at capturing the distinctions among the classes of the first category.

Our rejection of this judgement comes from believing that the goal of recovering class structure is as worthy as minimizing classification errors. The interpretability of the results of C_4 is superior to those of C_1 since it has better captured the nature of the underlying phenomenon. This means that the errors committed by C_4 are likely to be inside the same category of the correct response (given the nearly block diagonal structure of its heat map) while in the case of C_1, for example, the probability of having stimuli of the first category erroneously predicted as y_4 is very high.

The EMA and the NIT factor prove apt at considering the value of representing the underlying structure with their tight relation to perplexity. In fact, according to [23], while C_4, C_6 and C_7 focused on solving the so-called *domain adaptation problem*—the mismatch in training and testing conditions—with advanced machine learning techniques, many of the other teams, including C_1, addressed it by placing more weight on the labeled *test* samples provided along with the *train* samples, when validating the learned classifier, thus *explicitly* boosting test set accuracy.

Table 1. Perplexities, accuracy ($a(P_{XY})$), *EMA* ($a'(P_{XY})$) and *NIT factor* ($q(P_{XY})$) for MEG Mind Reading confusion matrices ranked by accuracy.

Exp.	$k_{X\|Y}$	μ_{XY}	$a(P_{XY})$	$a'(P_{XY})$	$q(P_{XY})$
C_1	2.562	1.932	0.680	0.390	0.386
C_2	2.447	2.023	0.632	0.409	0.405
C_3	2.589	1.912	0.628	0.386	0.382
C_4	2.430	2.037	0.622	0.412	0.407
C_5	2.723	1.818	0.565	0.367	0.364
C_6	2.682	1.846	0.542	0.373	0.369
C_7	2.730	1.813	0.539	0.366	0.363
C_8	3.629	1.364	0.472	0.276	0.273
C_9	2.995	1.653	0.443	0.334	0.331
C_{10}	4.801	1.031	0.242	0.208	0.206

Class C_4 should have been ranked above the rest by EMA or NIT factor (in all cases $k=5$ and $k_X=4.950$).
doi:10.1371/journal.pone.0084217.t001

Discussion

Measure definition

Perplexity has already been used as a performance measurement for language modeling where it refers to the expected average of alternatives a model has at every word history [24]. It is also often used as an off-line method for speech recognition task evaluation following the intuition that a classifier using a lower-perplexity model will outperform a higher-perplexity one, all other things equal.

It cannot be stressed enough that since the EMA and the NIT factor concentrate in the prior class distribution and mutual information, it is harder for classifiers to boost their performance by manipulating the posterior class distribution through specialization: only the increase in information transfer through $MI_{P_{XY}}$ will improve the evaluation figure.

Considering robustness, the EMA, being a harsher, worst-case criterion, might be more deserving of trust than easygoing and unreliable accuracy to, for instance, guide decision making. It certainly has a more interpretable and less easily bendable criterion—specially if *reporting* the classification error is not the

ultimate goal. Furthermore, in cases where $k \neq m$—for instance, when using a "reject" class—the EMA and the NIT factor are still defined, whereas accuracy is problematic, and not very much used.

Classification task assessment

As seen in the MEG mind reading example, the EMA and the NIT factor are capable of determining whether a task has been effectively solved or not. But it cannot distinguish whether this is caused by technical limitations in the classifier selection process or because the task is inherently "hard". Only the kind of iterated classification effort of community research that attempts many different classifier-building techniques on the same task can be effective for this purpose.

Nevertheless, the effective input perplexity k_X can ensure that, methodologically at least, the task is "as hard as it should be" at $k_X \approx k$. Furthermore, our developments show clearly that a failure to maintain prior class distribution uniformity in the design or capture of the task data entails that the expected mutual information—therefore the NIT factor —captured by any possible classifier that solves the task can never reach maximal levels. This is a strong guideline for prospective collectors of datasets, although data balancing strategies after data collection can also be used to achieve this goal [19].

Measure comparison

Several other measures have sprouted to deal with the inadequacies of accuracy such as the Area-Under-the-(ROC)-Curve [8,25], the Variation of Information [21], the Relative Classifier Information [26], the Confusion Entropy [10,27] or Cohen's Kappa [13], but their use is not widespread, specially for the non-binary case, due to complexity of calculation, disparate purposes or each measures' own shortcomings. For instance, the AUC first needs to find a (multiclass) ROC representation of the task by obtaining multiple classifiers, possibly with the help of a parameter in the classifier learning process. The trading for good-vs-wrong decisions in terms of the parameter can then be judged from the Area-Under-the-ROC curve, which is then a measure *on the learning method or model*. In contrast, EMA would provide a different point in the ET for each classifier whence the best of these classifiers could be chosen. Complementarily, on the *population of classifiers*, a statistical description of the NIT factor could be used to assess the learning capabilities of the method.

In classification proper, to illustrate the disparity of the conclusions that can be reached with alternative performance measures, we have included in *File S1* a comparison of the classical Matthew Correlation Coefficient (MCC) [28] and the Confusion Entropy (CEN) [27]—whose similarities are also explored in [10]–on three different classifications tasks: the MEG Mind Reading task already explored, the TASS sentiment analysis task [29]–both machine learning tasks—and the well-known Miller & Nicely human perceptual capability exploration task [5].

For each task we provide the heat maps of the confusion matrices (Figs. S1, S2 and S4 in *File S1*) as customary. We also provide the tables detailing perplexities, EMA, NIT factor, $1-\mathrm{CEN}$ and MCC' related values (Tables S1, S2 and S3 in *File S1*). The entries in the tables are ordered by accuracy. For the TASS and M&N data we also supply the ET's with the color bar according to EMA, $1-\mathrm{CEN}$ and MCC' (Figs. S3 and S5). Their comparison, detailed in the *File S1* Section, reveals that MCC' is highly correlated with accuracy in ranking results and shows similar shortcomings. Even though CEN performs a little better, it is highly biased towards majority classifiers providing over optimistic assessment for them. Notably, once the ET, EMA and the NIT factor have shed light on the problem, reassessment of prior evidences for either CEN or MCC prove them not to be so advantageous in evaluating classifiers.

Materials and Methods

The entropy triangle

Consider two discrete random variables X and Y and their joint probability distribution P_{XY}. An entropy diagram somewhat more complete than what is normally used for the relations between their entropies was presented in [12] and is here depicted in Fig. 5(A). We distinguish in it the familiar decomposition of the joint entropy $H_{P_{XY}}$ as the two entropies H_{P_X} and H_{P_Y} whose intersection is $\mathrm{MI}_{P_{XY}}$. But notice that the increment between $H_{P_{XY}}$ and $H_{P_X \cdot P_Y}$ is yet again $\mathrm{MI}_{P_{XY}}$, hence the expected mutual information appears *twice* in the diagram. Further, the interior of the outer rectangle represents $H_{U_X \cdot U_Y}$—with U_X and U_Y the uniform distribution on inputs and outputs—,the interior of the inner rectangle $H_{P_X \cdot P_Y}$, and $\Delta H_{P_X \cdot P_Y}$ is their difference. Finally, the *variation of information* $VI_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}}$ was found to be an important quantity in [21]. Putting together this information results in the *balance equation for information related to a joint distribution*,

$$H_{U_X \cdot U_Y} = \Delta H_{P_X \cdot P_Y} + 2MI_{P_{XY}} + VI_{P_{XY}}$$

which can be further normalized in $H_{U_X \cdot U_Y}$,

$$1 = \Delta H'_{P_X \cdot P_Y} + 2MI'_{P_{XY}} + VI'_{P_{XY}} \tag{6}$$

and represented in a De Finetti or ternary diagram as the equation of the 2-simplex in normalized $\Delta H'_{P_X \cdot P_Y} \times 2MI'_{P_{XY}} \times VI'_{P_{XY}}$ space, hence the name entropy triangle, *ET*.

The position of the coordinates of a classifier on the Entropy Triangle characterizes its performance, and we use this characterization to visually assess it indicated in Fig. 6. Classifiers at the apex or close to it obtain the highest accuracy possible on balanced datasets and transmit a lot of mutual information, hence they are the *best classifiers* possible. Those at the left vertex or close to it are dealing with balanced data but doing a bad job of utilizing it: they are the *worst classifiers*. Those at the right vertex or close to it are dealing with very easy, unbalanced data and claiming very high accuracy, yet they are not learning anything from it: they are *specialized (majority) classifier*s and our intuition is that they are the kind of classifiers that generate the accuracy paradox [16].

The split entropy triangle

Notice that in equation (6), since both U_X and U_Y and P_X and P_Y are independent as marginals of $U_X \cdot U_Y$ and $P_X \cdot P_Y$, respectively, we may write:

$$\Delta H_{P_X P_Y} = (H_{U_X} - H_{P_X}) + (H_{U_Y} - H_{P_Y})$$

$$= \Delta H_{P_X} + \Delta H_{P_Y},$$

what suggests writing separate balance equations for each variable,

$$H_{U_X} = \Delta H_{P_X} + \mathrm{MI}_{P_{XY}} + H_{P_{X|Y}}$$

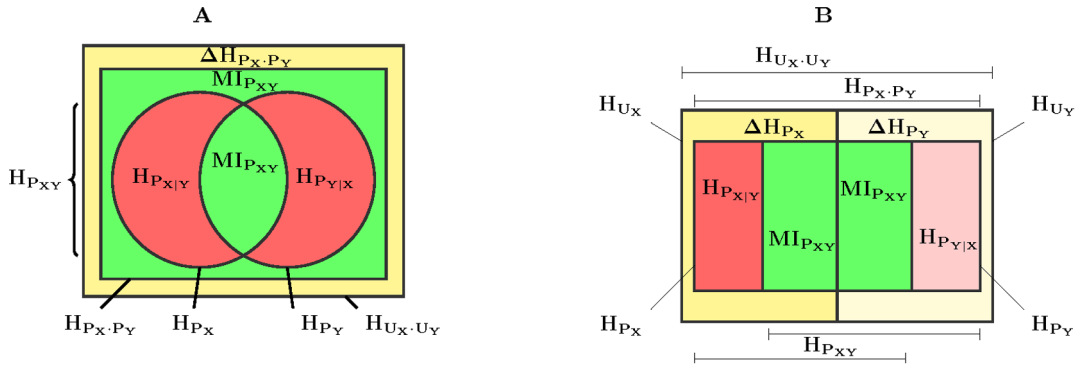

Figure 5. (Color online) Extended information diagrams of entropies related to a bivariate distribution: (A) conventional diagram, and (B) split diagram. The bounding rectangle is the joint entropy of two uniform (thence independent) distributions U_X and U_Y of the same cardinality as P_X and P_Y. The expected mutual information $MI_{P_{XY}}$ appears *twice* in (A) and this makes the diagram split for each variable symmetrically in (B).
doi:10.1371/journal.pone.0084217.g005

$$H_{U_Y} = \Delta H_{P_Y} + \mathrm{MI}_{P_{XY}} + H_{P_{Y|X}}\,.$$

The formulae above and the occurrence of twice the expected mutual information in equation (6) suggests a different information diagram, depicted in Fig. 5(b): both variables X and Y now appear somehow decoupled—in the sense that the areas representing them are disjoint—yet there is a strong coupling in that the expected mutual information appears in both H_{P_X} and H_{P_Y}. It is important to note that both decompositions can be represented in the same (split) entropy triangle as equation (6) dictates. The technique is explained in [12].

Data

The space of $k \times k$ square confusion matrices, C_{XY} of sizes $k \in \{2,3,4\}$ and a given number of input samples, N, depicted in Fig. 2 was obtained by first generating every possible partition of N with k parts as input distributions P_X, allocating $n_i, i \in \{1,2,\ldots,k\}$ input samples in each of the input classes. In this way, the set of all possible input class distributions, from uniform U_X to the most skewed P_X, is obtained. Then, for each of the previous distributions, every possible weak composition of n_i with k parts is produced, yielding k sets of all the possible distributions for each of the rows of C_{XY}. Finally, the Cartesian product of those sets produces every possible combination of rows corresponding to the selection of one element in every one of the sets. Except from row permutations —that would only amount to a reordering of the input classes— this procedure guarantees the presence of every possible C_{XY}.

The MEG mind reading task aims at decoding the identity of a video stimulus based on magnetoencephalography (MEG) recordings done during naturalistic stimulation [23]. In particular, subjects were exposed to video stimuli of different classes: a first category of *short* clips (6–26 s. long) with x_1 being *artificial* stimuli (screen savers showing animated shapes or text), x_2 being *natural* stimuli (sceneries like mountains or oceans) and x_3 being *football* stimuli (from —European— football matches) and a second

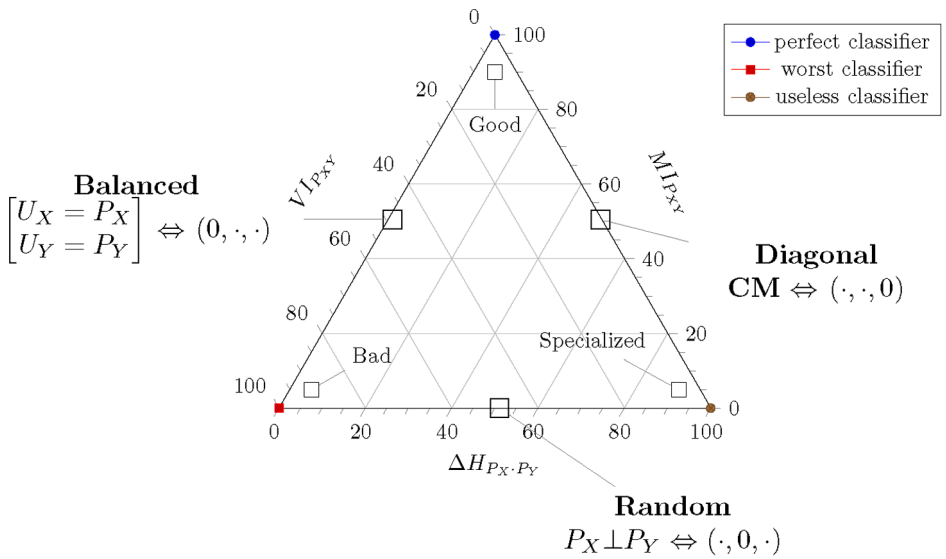

Figure 6. Schematic Entropy Triangle showing interpretable zones and extreme cases of classifiers. The annotations on the center of each side are meant to hold for that whole side.
doi:10.1371/journal.pone.0084217.g006

category of *long* clips (approximately 10 min. long) with x_4 being television series (from "Mr. Bean" in particular) and x_5 being films (from Chaplin's "Modern times"). The goal was to classify unlabeled test examples into these classes based on the MEG signal alone. The competition took place in March, 2011 and 10 participants submitted their classifiers whose confusion matrices are analyzed in this paper. The data was provided upon request from the organizers of the competition.

The MATLAB(A registered trademark of The MathWorks, Inc.) code to draw the entropy triangles in Figures 2 and 4 has been made available at: http://www.mathworks.com/matlabcentral/fileexchange/30914

Supporting Information

Figure S1 Heat maps of the classifiers of the MEG mind reading competition [23]. Rows correspond to stimulus $X = x_i$ and columns to the decision $Y = y_j$ or response. Darker hues correlate with higher joint probability P_{XY}. The classifier denominations obey to their position in the ranking produced by accuracy.
(TIFF)

Figure S2 Heat maps of the classifiers of the TASS competition [29]. Rows correspond to stimulus $X = x_i$ and columns to the decision $Y = y_j$ or response. Darker hues correlate with higher joint probability P_{XY}. The classifier denominations obey to their position in the ranking produced by accuracy **A** Color bar represents EMA **B** Color bar represents $1-\mathbf{CEN}$ **C** Color bar represents $\mathbf{MCC}'$.
(TIFF)

Figure S3 (Color online) **Entropy decomposition for the classifiers of the TASS competition (A) with the color bar representing EMA, (B) $1-\mathbf{CEN}$, and (C) $\mathbf{MCC}' = (\mathbf{MCC}+1)/2$.**
(TIFF)

Figure S4 Heatmaps of the classifiers of the TASS competition [29]. Rows correspond to stimulus $X = x_i$ and columns to the decision $Y = y_j$ or response. Darker hues correlate with higher joint probability P_{XY}. The classifier denominations obey to their position in the ranking produced by accuracy **A** Color bar represents EMA **B** Color bar represents $1-\mathbf{CEN}$] withFigures **C** Color bar represents $\mathbf{MCC}'$.
(TIFF)

Figure S5 (Color online) **Entropy decomposition for MN phonetic confusion matrices (A) with the color bar representing EMA, (B) $1-\mathbf{CEN}$, and (C) $\mathbf{MCC}' = (\mathbf{MCC}+1)/2$.**
(TIFF)

File S1 Supporting Information. A comparison of the classical Matthew Correlation Coefficient (MCC) [28] and the Confusion Entropy (CEN) [27]—whose similarities are also explored in [10]–on three different classifications tasks: the MEG Mind Reading task already explored, the TASS sentiment analysis task [29]–both machine learning tasks—and the well-known Miller & Nicely human perceptual capability exploration task [5].
(PDF)

Acknowledgments

The authors would like to thank A. Klami for providing the MEG Mind Reading data, J. Villena for the TASS data, and both A. Sánchez, C. Bousoño and the anonymous reviewers for comments on previous versions of this paper.

Author Contributions

Conceived and designed the experiments: FJVA CPM. Performed the experiments: FJVA CPM. Analyzed the data: FJVA CPM. Contributed reagents/materials/analysis tools: FJVA CPM. Wrote the paper: FJVA CPM.

References

1. Sokal RR (1974) Classification: Purposes, principles, progress, prospects. Science 185: 1115–1123.
2. Huang H, Liu CC, Zhou XJ (2010) Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. Proceedings of the National Academy of Sciences of the United States of America 107: 6823–6828.
3. West M, Blanchette C, Dressman H, Huang E, Ishida S, et al. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. Proceedings of the National Academy of Sciences of the United States of America 98: 11462–11467.
4. Wei X, Li KC (2010) Exploring the within- and between-class correlation distributions for tumor classification. Proceedings of the National Academy of Sciences of the United States of America 107: 6737–6742.
5. Miller GA, Nicely PE (1955) An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America 27: 338–352.
6. Congalton RG, Green K (1999) Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. CRC Press, Inc.
7. Jurafsky D, Martin JH (2000) Speech and Language Processing. Prentice-Hall.
8. Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 240: 1285–1293.
9. Powers D (2011) Evaluation: From Precision, Recall and F-Measure to ROC., Informedness, Markedness & Correlation. Journal of Machine Learning Technologies 2: 37–63.
10. Jurman G, Riccadonna S, Furlanello C (2012) A comparison of MCC and CEN error measures in multi-class prediction. PLoS ONE 7.
11. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27: 861–874.
12. Valverde-Albacete FJ, Peláez-Moreno C (2010) Two information-theoretic tools to assess the performance of multi-class classifiers. Pattern Recognition Letters 31: 1665–1671.
13. Ben-David A (2007) A lot of randomness is hiding in accuracy. Engineering Applications of Artificial Intelligence 20: 875—885.
14. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Information Processing & Management 45: 427–437.
15. Kononenko I, Bratko I (1991) Information-based evaluation criterion for classifier's performance. Machine Learning 6: 67–80.
16. Zhu X, Davidson I (2007) Knowledge discovery and data mining: challenges and realities. Premier reference source. Information Science Reference.
17. Thomas C, Balakrishnan N (2008) Improvement in minority attack detection with skewness in network traffic. Proc SPIE Int Soc Opt Eng 6973: 69730N-69730N-12.
18. Fernandes JA, Irigoien X, Goikoetxea N, Lozano JA, naki Inza I, et al. (2010) Fish recruitment prediction, using robust supervised classification methods. Ecological Modelling 221: 338–352.
19. Garcia-Moral A, Solera-Urena R, Pelaez-Moreno C, Diaz-de Maria F (2011) Data balancing for efficient training of hybrid ann/hmm automatic speech recognition systems. Audio, Speech, and Language Processing, IEEE Transactions on 19: 468–481.
20. Fano RM (1961) Transmission of Information: A Statistical Theory of Communication. The MIT Press, 400 pp.
21. Meila M (2007) Comparing clusterings|an information based distance. Journal of Multivariate Analysis 28: 875–893.
22. Jaynes E (1983) Concentration of distributions at entropy maxima. In: Rosenkrantz R, editor, Papers on Probability, Statistics, and Statistical Physics, D. Reidel Publishing.
23. Klami A, Ramkumar P, Virtanen S, Parkkonen L, Hari R, et al. (2011) ICANN/PASCAL2 challenge: MEG mind reading – overview and results. In: Klami A, editor, Proceedings of ICANN/PASCAL2 Challenge: MEG Mind Reading. Espoo, Aalto University Publication series SCIENCE + TECHNOLOGY 29/2011, pp. 3–19.
24. Jelinek F (1997) Statistical Methods for Speech Recognition. Cambridge, Ma; London, UK: The MIT Press.
25. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30: 1145–1159.
26. Sindhwani V, Bhattacharya P, Rakshit S (2001) Information theoretic feature crediting in multiclass support vector machines. In: Proceedings of the First SIAM International Conference on Data Mining. pp.5–7.

27. Wei JM, Yuan XJ, Hu QH, Wang SQ (2010) A novel measure for evaluating classifiers. Expert Systems with Applications 37: 3799–3809.
28. Matthews B (1975) Comparison of the predicted and observed secondary structure of {T4} phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure 405: 442–451.
29. Valverde-Albacete FJ, Carrillo-de Albornoz J, Peláez-Moreno C (2013) A proposal for new evaluation metrics and result visualization technique for sentiment analysis tasks. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Springer Berlin Heidelberg, volume 8138 of *Lecture Notes in Computer Science*. pp. 41–52.